On prediction and density estimation

Peter McCullagh
University of Chicago
December 2004

Summary

Having observed the initial segment of a random sequence, subsequent values may be predicted by calculating the conditional distribution given what has been observed. In statistical applications, it is usually necesary to work with a parametric family of processes, so the predictive distribution depends on the parameter, which must be estimated from the data. This device is used in a finite-dimensional parametric model, so that maximum-likelihood can be applied to density estimation. An exchangeable process is constructed by generating a random probability distribution having a smooth density, and subsequently generating independent and identically distributed components from that distribution. The random probability distribution is determined by the squared modulus of a complex Gaussian process,and the finite-dimensional joint densities of the resulting process are obtained in the form of matrix permanents. The conditional density of Y_{n+1} given $(y_1, \ldots, y_n)$ is obtained as a weighted baseline, in which the permanent is the weight function. For prediction in the sense of density estimation, the permanent plays a role similar to that of the likelihood function in parametric inference.

Some key words: Complex Gaussian process; spatial-temporal Cox process; exchangeable process; matrix permanent; permanent density estimator;

1. Introduction

The aim of this paper is to construct a parametric statistical model suitable for density estimation. Given the number of articles on the subject of bandwidth selection, and the variety of algorithms available for computation of kernel smoothers, it may seem quixotic to inquire into the goal of density estimation. Nevertheless, that question must come first. The most natural answer is to identify density estimation with prediction for an exchangeable process. The standard interpretation emphasizing estimation and mean-squared error (Silverman 1986, section 1.1), leads to a mathematically different formulation of the problem. But the goals are sufficiently hard to distinguish that we are inclined to regard prediction and density estimation as the same activity.

The chief goal is to construct a statistical model suitable for density estimation in the sense that predictive distributions resemble kernel density estimators. The key step is to construct a suitable random probability distribution, which is done using a complex-valued Gaussian process with covariance function K. The finite-dimensional joint distributions of the resulting process are obtained in terms of matrix permanents. The predictive density, or density estimate, or conditional density given the observation, is the baseline density weighted by a ratio of permanents.

The model has one principal element that may be chosen more or less arbitrarily. Smoothness of the Gaussian process is governed primarily by the behaviour of K near the origin. A covariance function such as $K(y, y') = \exp(-|y - y'|^2)$ is ordinarily not recommended for spatial processes because its realizations are smooth functions. Stein (1999) shows that they are mean square analytic. But this very property makes it suitable for density estimation where smoothness carries a premium and roughness is penalized.

2. Density estimation

2.1 *Exchangeable sequences*

Let $Y = (Y_1, Y_2, \ldots)$ be an infinitely exchangeable real-valued process with finite-dimensional distributions P_n defined on $\mathcal{R}^n$. In connection with density estimation, we assume that P_n has a density p_n. An observation $y = (y_1, \ldots, y_n)$ consisting of n components, determines a conditional distribution, or predictive distribution for Y_{n+1}, whose density at t is

$$h_n(t; y_1, \ldots, y_n) = \frac{p_{n+1}(y_1, \ldots, y_n, t)}{p_n(y_1, \ldots, y_n)}.$$

The function h that associates with each n and with each point $y \in \mathcal{R}^n$ the distribution on $\mathcal{R}$ with density $h_n(\cdot; y)$ is called the density estimator associated with the process. Every exchangeable process having finite-dimensional density functions is thus associated with a density estimator.

By the de Finetti characterization, an exchangeable process is a mixture of independent and identically distributed processes. Thus, the process may be generated by first selecting a distribution F at random from the set of distributions on the real line. Subsequently, the sequence is generated by selecting $Y_i \sim F$, independently for each component. If F were given, the conditional distribution or optimal predictor for Y_{n+1} would simply be F, in one sense the distribution by which the process was generated. If F is not given, the best predictor may be regarded as the best available estimator of F. Hence, for an exchangeable sequence, there appears to be little practical difference between density estimation and optimal prediction.

2.2 *Statistical model*

To construct a statistical model for an exchangeable real-valued sequence it is sufficient to construct a probability distribution Q on the set $\mathcal{P}(\mathcal{R})$ of probability distributions on the real line. A random element F with distribution Q is by definition a probability distribution on $\mathcal{R}$. For present purposes, we aim to construct Q so that the random element F has a smooth density with high probability. The construction proceeds as follows.

Let p_0 be a given baseline density function, and let Z be a zero-mean complex-valued stationary Gaussian process on the real line with covariance function $K(y, y')$ depending only on $|y - y'|$. Then the squared modulus $|Z|^2$ is a non-negative real-valued process, and

$$\frac{|Z(y)|^2 p_0(y)}{\int |Z(t)|^2 \, p_0(t) \, dt} \tag{2.1}$$

is a weighted probability density on the real line. The smoothness of the Gaussian process is governed by the behaviour of K near the diagonal, so we may choose K to achieve the desired degree of smoothness in the densities generated.

If F is chosen according to (2.1), the joint conditional density at $(y_1, \ldots, y_n)$ is the product

$$\frac{|Z(y_1)|^2 \cdots |Z(y_n)|^2 \, p_0(y_1) \cdots p_0(y_n)}{\big(\int |Z(t)|^2 \, p_0(t) \, dt\big)^n}$$

and the n-dimensional distribution for the process is the expected value of this ratio. In section 3, we show how the problem may be reformulated as a Cox process in such a way that the problem of calculating the expectation of the ratio is avoided. For the moment we adopt the simpler strategy of approximating the expectation of the ratio by the ratio of expectations.

For a zero-mean Gaussian process, it is shown in the Appendix that

$$E\big(|Z(y_1)|^2 \cdots |Z(y_n)|^2\big) = \mathrm{per}_n[K](y_1, \ldots, y_n)$$

where $\mathrm{per}_n[K](y)$ is the permanent of the $n \times n$ matrix with entries $K(y_i, y_j)$:

$$\mathrm{per}_n[K](y) = \sum_\pi K(y_1, y_{\pi(1)})K(y_2, y_{\pi(2)}) \cdots K(y_n, y_{\pi(n)})$$

with summation over the $n!$ permutations. The approximation in the preceding paragraph implies that the expected value of the ratio is the ratio of expected values, so the n-dimensional joint density is proportional to the product

$$\mathrm{per}_n[K](y)\, p_0(y_1) \cdots p_0(y_n). \tag{2.2}$$

The conditional density at t for Y_{n+1} given $(y_1, \ldots, y_n)$ is thus

$$h_n(t; y) \propto \frac{\mathrm{per}_{n+1}[K](y_1, \ldots, y_n, t)\, p_0(t)}{\mathrm{per}_n[K](y_1, \ldots, y_n)}. \tag{2.3}$$

The predictive density is shown in Fig. 1 for four sample configurations.

The more elaborate construction in section 3, which avoids the simplifying assumption of the preceding paragraph, produces a similar expression for the conditional density with K replaced by a modified covariance function. The approximation has no effect on the form of the conclusion. The key result that the joint densities are determined by matrix permanents carries over without modificaion to bivariate or multivariate processes, or to processes on the unit circle or sphere. In this sense, density estimation in higher dimensions is not fundamentally more difficult than density estimation in one dimension.

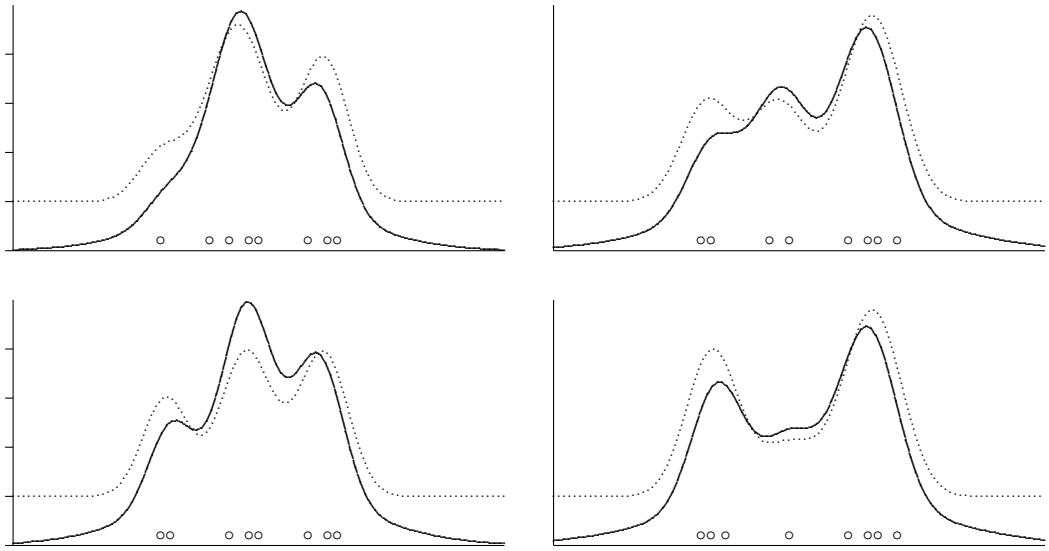

Fig. 1: Predictive density estimate (solid lines) for four sample configurations using a gaussian baseline with $\sigma = 1.4s$. Permanent ratios $\mathrm{per}_{n+1}[K](y,t)/\,\mathrm{per}_n[K](y)$ are shown as dotted lines. In each case $K(y, y') = \exp(-|y - y'|^2/\lambda^2)$ with λ equal to one quarter of the range of y.

In the argument leading to this point, the covariance function and the baseline density may be chosen arbitrarily. The range of the covariance function plays a role similar to that of the bandwidth in kernel density estimation, which means that it has a substantial effect on the shape of the predictive distribution. The bandwidth analogy must not be pushed too far because the process need not be ergodic, and K need not tend to zero at large distances. The predictive distributions in Figure 1 were obtained using a Gaussian covariance function $\exp(-|x-x'|^2/\lambda^2)$ with range parameter λ equal to one quarter of the range of the sample values. The baseline density is Gaussian with mean $\bar{x}$ and standard deviation $1.4s$. These values were chosen to illustrate the variation in shape of the predictive distribution that can be achieved by a three-parameter model.

3. Cox process

3.1 *Spatial-temporal representation*

Consider a spatial-temporal Cox process (Cox, 1955) for which the intensity at (y,t) is $|Z(y)|^2\,p_0(y)$, constant in time and spatially integrable but non-homogeneous. The temporal intensity is $\|Z\|^2 = \int |Z(y)|^2\,p_0(y)\,dy$, so the process evolves at a constant random rate. Conditionally on Z, the number of events in $(0,t)$ is Poisson with parameter $\|Z\|^2 t$.

Let T_n be the time taken to observe n events. The conditional distribution of T_n is Gamma with mean $n/\|Z\|^2$ and index n. The joint density at (t,y) of the time and the values is

$$\frac{\exp(-t\|Z\|^2)(\|Z\|^2)^n\,t^{n-1}}{\Gamma(n)}\,\frac{|Z(y_1)|^2\cdots|Z(y_n)|^2\,p_0(y_1)\cdots p_0(y_n)}{(\|Z\|^2)^n}$$

which reduces to

$$\exp(-t\|Z\|^2)t^{n-1}|Z(y_1)|^2\cdots|Z(y_n)|^2\,p_0(y_1)\cdots p_0(y_n)/\Gamma(n).$$

To compute the unconditional joint distribution we need to calculate the mean of this product of complex Gaussian variables.

It is mathematically convenient in what follows to assume that the space of observations is the unit circle in the complex plane rather than the real line. A particular point is denoted by $w = e^{i\theta}$ with $-\pi < \theta \le \pi$. On this space, we take p_0 to be the uniform distribution, and we take K to be invariant under rotation. Apart from the constant function, the eigenfunctions of K come in pairs $w^r, \bar{w}^r$ with eigenvalue $\lambda_r = \lambda_{-r}$ so that

$$K(w,w') \equiv K(w/w') = \sum_{r=-\infty}^{\infty} \lambda_r (w/w')^r.$$

Let $t \ge 0$ be a positive constant and let K_t be the stationary covariance function with eigenvalues $\lambda_r/(1+t\lambda_r)$. In other words,

$$K_t(w) = \sum w^r \lambda_r/(1+t\lambda_r),$$

so that $K_0 = K$, and the limit as $t \to \infty$ is such that tK_t has a flat spectrum corresponding to white noise. It is assumed that $\sum \lambda_r < \infty$, so the determinant ratio

$$D(t) = \prod_{r=-\infty}^{\infty} (1+t\lambda_r)$$

is finite.

The joint density at (t,w) of the time and the values given Z is

$$\exp(-t\|Z\|^2)t^{n-1}|Z(w_1)|^2\cdots|Z(w_n)|^2 \;/\; \big((2\pi)^n\,\Gamma(n)\big).$$

where the points $w_1,\ldots,w_n$ are now on the unit circle, From the Appendix, the unconditional joint density of the values and the time T_n is

$$\frac{t^{n-1}\ \mathrm{per}_n[K_t](w_1,\ldots,w_n)}{(2\pi)^n\,D(t)\,\Gamma(n)} \tag{3.1}$$

with K replaced by K_t. Since T is not observed, the final step is to calculate the marginal density of the values by integration. Numerical integration and Laplace approximation are both feasible options, but the recommended alternative is to regard the accrual rate $\rho = n/T_n$ as a parameter to be estimated. The conditional distribution given ρ is proportional to $\mathrm{per}_n[K_{n/\rho}](w_1,\ldots,w_n)$ from which ρ can in principle be estimated by maximum likelihood. The maximum likelihood predictive density at w is then proportional to

$$\mathrm{per}_{n+1}[K_{(n+1)/\hat{\rho}}](w_1,\ldots,w_n,w). \tag{3.2}$$

In practice, unless K is fully specified, there may be additional parameters to be estimated. Nonetheless, the predictive density has the same form with K_t replaced by a suitable estimate $\hat{K}_t$.

3.2 *Specific details*

The choice of covariance function K and the method of parameter estimation both affect the predictive density. Details of the choices made in Fig. 2 are now given, together with the invertible transformation used to map the real line onto the unit circle.

On the unit circle, we use the Poisson covariance function

$$K(e^{i\theta}) = e^{-\alpha}\Re\sum_0^\infty \alpha^r w^r/r! = e^{-\alpha}e^{\alpha\cos\theta}\cos(\alpha\sin\theta)$$

with $w = e^{i\theta}$ and α^{-1} as the angular range parameter. In principle, α could be estimated from the data by maximum likelihood, but the information available is very small, so the plots in Fig. 2 use two values $\alpha = 0.5, 1.0$ to illustrate the effect on the predictive density. The covariance function is infinitely differentiable at $\theta = 0$, which forces the Z-process to be smooth as required. There is no loss of generality in taking $K(1) = 1$. The eigenvalues are the Poisson probabilities $\lambda_0 = e^{-\alpha}$ and $2\lambda_r = e^{-\alpha}\alpha^r/r!$ for $r > 0$.

Maximum likelihood is not feasible, so the parameter t is estimated by maximizing the function

$$t^n \ \mathrm{per}_n[K_t](w_1, \ldots, w_n)/D(t) \tag{3.3}$$

This is not a likelihood function, but maximization is the step required to compute the integral of (3.1) with respect to t by Laplace approximation.

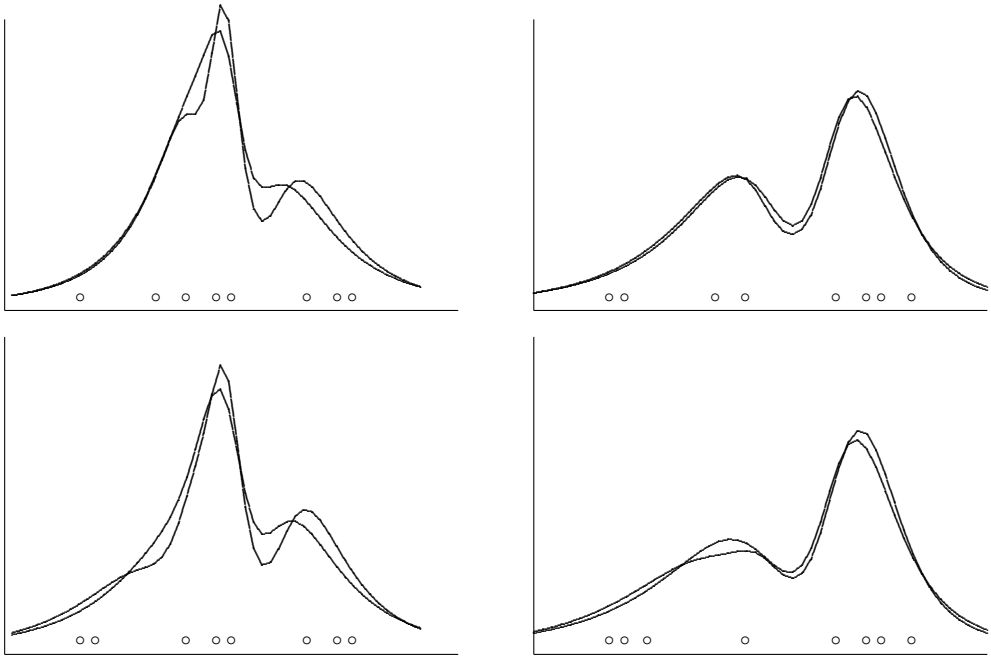

Fig. 2: Predictive density estimates for four sample configurations using a Cauchy baseline. The range parameter of the covariance function affects the smoothness, smaller α giving smoother densities. The smoother density corresponds to a range parameter $\alpha = 0.5$.

Finally points on the real line are transformed to the unit circle by the transformation

$$w_j = -\frac{y_j - \mu - i\sigma}{y_j - \mu + i\sigma}$$

with (μ, σ) chosen to make $\sum w_j = 0$. In other words, (μ, σ) is the maximum-likelihood estimator of the parameter in the Cauchy location-scale model. In principle, all four parameters (μ, σ, α, t) could be estimated simultaneously, but this was not done for the graphs in Fig. 2.

4. Stochastic approximation

The model in section 2 uses a zero-mean Gaussian process as a weight function to modulate the baseline density. We now consider a similar model in which Z is replaced by $\tau + Z$ where τ is a complex number and Z is a zero-mean stationary Gaussian process. The probability density (2.1) becomes

$$\frac{|\tau + Z(y)|^2 p_0(y)}{\|\tau + Z\|^2} = \frac{1 + \epsilon(y) + \bar{\epsilon}(y) + |\epsilon(y)|^2}{1 + \|\epsilon\|^2} p_0(y),$$

where $\epsilon(y) = Z(y)/\tau$ and $\|\epsilon\|^2 = \int |\epsilon^2(y)|^2 \, p_0(y) \, dy$.

For large τ, the stochastic expansion of the joint conditional density is moderately complicated, but many of the terms such as $\epsilon(y)$ and $\epsilon(y)\epsilon(y')$ have zero expectation and thus do not contribute to the unconditional joint density. The terms having non-zero expectation in the expansion up to second order in τ are

$$1 - n\|\epsilon\|^2 + \sum_{jk} \epsilon(y_j)\bar{\epsilon}(y_k).$$

To this order of approximation, the unconditional joint density at $(y_1, \ldots, y_n)$ is

$$\exp\Big(\tau^{-2} \sum_{jk} K(y_j, y_k) - n\sigma^2/\tau^2\Big) \times p_0(y_1) \cdots p_0(y_n)$$

where $\sigma^2 = K(y, y)$ is the variance of the process. The conditional density at y given the observed values is proportional to

$$p_0(y) \times \exp\Big(\tau^{-2} \sum_j (K(y, y_j) + K(y_j, y))\Big).$$

This approximation resembles a kernel density estimator in the sense that it is an additive function of the observed components $(y_1, \ldots, y_n)$. For fixed n, the error of this approximation tends to zero as $\tau \to \infty$: it is not an asymptotic approximation for large samples.

5. Concluding remarks

The goal of the paper was to construct an exchangeable process whose predictive densities resemble the output of a kernel density estimator. As the graphs demonstrate, the processes have predictive densities that are smooth, and potentially multi-modal depending on the configuration of points observed. In that sense, they resemble kernel density estimators. Mathematically, the density estimator (3.2) differs from a kernel density estimator in one crucial respect: it is not an additive function of the observed components. In that sense, perhaps, the resemblance may be superficial, but the analysis in section 4 suggests that additive approximations may be available.

For the processes considered in section 3, the ratio of the predictive density to the baseline is proportional to the permanent $\mathrm{per}_{n+1}[K_t](y_1, \ldots, y_n, y)$. For these processes, the permanent plays a role in prediction similar to that of the likelihood in parametric inference. Unfortunately, the numerical difficulty in computing the permanent (Valiant, 1979) means that it is difficult to translate this interpretation into a workable algorithm. However, the existence of a polynomial-time approximation algorithm (Jerrum, Sinclair and Vigoda 2003) suggests that the numerical issues may be surmountable.

Appendix: Complex Gaussian moments

Let $(Z_1, \ldots, Z_n)$ be a zero-mean random variable whose distribution is n-dimensional complex Gaussian with covariances $\mathrm{cov}(Z_i, Z_j) = 0$ and $\mathrm{cov}(Z_i, \bar{Z}_j) = K(i,j)$. The joint density at z with respect to Lebesgue measure in $\mathcal{R}^{2n}$ is $\pi^{-n}|K|^{-1}\exp(-z^*K^{-1}z)$, where K is Hermitian. In particular, if K is symmetric, i.e. real, the joint distribution on $\mathcal{R}^{2n}$ of the real and imaginary parts of Z is Gaussian with covariance matrix $K \otimes I_2/2$.

Let $X_j = |Z_j|^2$ be the squared modulus of the jth component. The expected value of the product $X_1 \cdots X_n$ is

$$E(X_1 \cdots X_n) = E(Z_1 \cdots Z_n \, \bar{Z}_1 \cdots \bar{Z}_n).$$

Since the Zs are zero-mean Gaussian and $E(Z_i Z_j) = 0$, the standard expression for moments in terms of cumulants (McCullagh, 1987) involves only partitions into blocks of size two, each block having one Z_i and one conjugated component $\bar{Z}_j$. Thus

$$\begin{aligned} E(X_1 \cdots X_n) &= \sum_\pi E(Z_1 \bar{Z}_{\pi(1)}) \cdots E(Z_n \bar{Z}_{\pi(n)}) \\ &= \sum_\pi K(1,\pi(1))\, K(2,\pi(2)) \cdots K(n,\pi(n)), \end{aligned} \tag{A1}$$

which is the permanent of the Hermitian matrix K. The joint cumulant of order n is also a sum of products, but the sum is restricted to the $(n-1)!$ cyclic permutations.

Now write (A1) as an integral over $\mathcal{R}^m$ with $m \geq n$

$$\int_{\mathcal{C}^m} |z_{j_1}|^2 \cdots |z_{j_n}|^2 \exp(-z^*K^{-1}z)\, dz = \pi^m |K| \times \mathrm{per}_n[K](j_1, \ldots, j_n),$$

where $|K|$ is the determinant of the $m \times m$ matrix, and $[K](j_1, \ldots, j_n)$ is the $n \times n$ matrix consisting of the designated rows and columns, possibly with duplicates. The expected value of the product $|Z_{j_1}|^2 \cdots |Z_{j_n}|^2 \exp(-Z^*QZ)$ may be obtained directly from the integral

$$\pi^{-m}|K|^{-1} \int_{\mathcal{C}^m} |z_{j_1}|^2 \cdots |z_{j_n}|^2 \exp(-z^*(K^{-1}+Q)z)\, dz$$

which simplifies to

$$|K|^{-1}|K^{-1}+Q|^{-1}\, \mathrm{per}_n[(K^{-1}+Q)^{-1}](j_1, \ldots, j_n) = |I+KQ|^{-1}\, \mathrm{per}_n[(K^{-1}+Q)^{-1}](j_1, \ldots, j_n).$$

If each eigenvector of K is also an eigenvector of Q, K with eigenvalues λ_r and Q with eigenvalues τ_r, the matrix $(K^{-1}+Q)^{-1}$ has eigenvalues $\lambda_r/(1+\lambda_r\tau_r)$. In the applications considered in the paper, Q is a multiple of the identity, $\tau = t$ is constant, and K_t is the matrix whose eigenvectors are those of K and whose eigenvalues are reduced to $\lambda_r/(1+t\lambda_r)$. The determinant factor is $|I+KQ| = \prod(1+t\lambda_r)$.

References

Cox, D.R. (1955) Some statistical methods connected with series of events (with discussion). *J. Roy Statist Soc. B* 27, 129–164.

Jerrum, M., Sinclair, A. and Vigoda, E. (2003) A polynomial-time approximation algorithm for the permanent of a matrix with non-negative entries. *J. Assoc. Comp. Mach.* (to appear).

McCullagh, P. (1987) *Tensor Methods in Statistics.* London, Chapman and Hall.

Silverman, B. (1986) *Density Estimation.* Chapman and Hall, London.

Stein, M. (1999) *Interpolation of Spatial Data.* Springer

Valiant, L. (1979) The complexity of computing the permanent. *Theoretical Computer Science* 8, 189–201.